Population Projection and Growth Parameter Estimation using LSTM Framework

Pandrangi Aditya Sriram[1]

***Abstract*— This paper attempts to project population and estimate various parameters of population growth using a predictive framework forecasting. Utilizing demographic data from the UN Dataset, the study utilizes various techniques on historical records to generate fine-grained temporal data. Utilizing this data, the model predicts future birth rates, death rates, migration rates, and population growth for a specific country, with Russia and Nigeria serving as case studies. A stacked LSTM architecture was employed to model and predict temporal dependencies in demographic trends using a time-series-based approach. The model iteratively projects population changes until 2100, demonstrating the integration of machine learning techniques into demographic forecasting. This study highlights the potential of neural networks to address complex time series forecasting challenges in population studies.**

I. Introduction

Forecasting population growth is a basic necessity in order to forecast urbanization trends, socioeconomic parameters, and migration patterns using historical and socioeconomic data, which will help predict demographic shifts, aiding policymakers, economists, and businesses in making data-driven decisions, such as in the fields of resource allocation and strategic (human as well as capital/monetary) investments. Population projection is vital for governments and planners due to their implications for pensions, healthcare, education, and other public services. Although projections come with inherent uncertainties, they are indispensable for informed decision making. Projections are often refreshed regularly by various organizations, including the United Nations Organization, to incorporate new data and trends, and alternative scenarios are developed probabilistically to explore different possible futures.

Key challenges and areas of improvements in population projection methodologies include:

1) Learnings from Population Forecasts: This involves learning from past projection errors to improve future forecasts.
2) Probabilistic Methods: These methods explicitly quantify uncertainty by estimating a range of possible outcomes rather than a single deterministic forecast.
3) Migration Projections: Predicting migration remains a significant challenge due to the unpredictability of international and internal movement patterns. This often has to do with the socioeconomic and political scenario of the country in question.
4) Expanding Projection Dimensions: Moving beyond the traditional focus on age, sex, and region to include socio-economic factors such as education, ethnicity, GDP Per Capita, and household structure.

Prior work used historically utilized simple deterministic models were utilized for population projections. However, today, sophisticated cohort component models are utilized by the United Nations, with algorithms often kept out of public reach. More methodologies shall be discussed in the Literature Review Section. Advancements in computational power and statistical techniques have allowed for more reliable analyses for even complex factors such as migration rate. One such methodology utilized in this paper is the use of LSTM, which follows the principle of ”past is future” - certain economic factors may cause a similar trend in the increase/decrease of migration rate as it caused in the past.

Although population projections will never be perfectly accurate, they can be improved by better data input, methodological innovations, and the integration of diverse research disciplines. Probabilistic approaches, in particular, offer a promising avenue for explicitly addressing and communicating the uncertainties inherent in demographic forecasting.

Against this backdrop, this paper attempts to utilize a deterministic regression-based LSTM approach with time-series modelling to estimate various parameters.

The projection of the population depends on the following three key parameters:

1) Birth Rate (per 1,000 people)
2) Mortality Rate (per 1,000 people)
3) Net Migration Rate (per 1,000 people)

The Birth Rate is defined as the number of live births per 1,000 people per year, while Mortality Rate is the number of deaths per 1,000 people per year. More sophisticated methods utilize autoencoders to find a correlation between birth and death rate, with a time gap of life expectancy in between. More complexity can be introduced via gender and socio-economic factors as described above.

The Net Migration Rate is defined in an inward direction - that is, from the other countries to the country in question. Thus, those migrating outwards would be counted as a net negative in the Migration Rate.

II. Literature Review

A number of algorithms have been reviewed and a suitable one chosen keeping while striking a balance between simplicity and accuracy/error of the model.

Chen, et al explored mortality rate forecasting using machine learning models such as LSTM, Bi-LSTM, and GRU, comparing them against the traditional models. This paper utilized U.S. mortality data from 1966 to 2015, the study found that LSTM with certain parameters, similar to those used in this paper, provided the most accurate

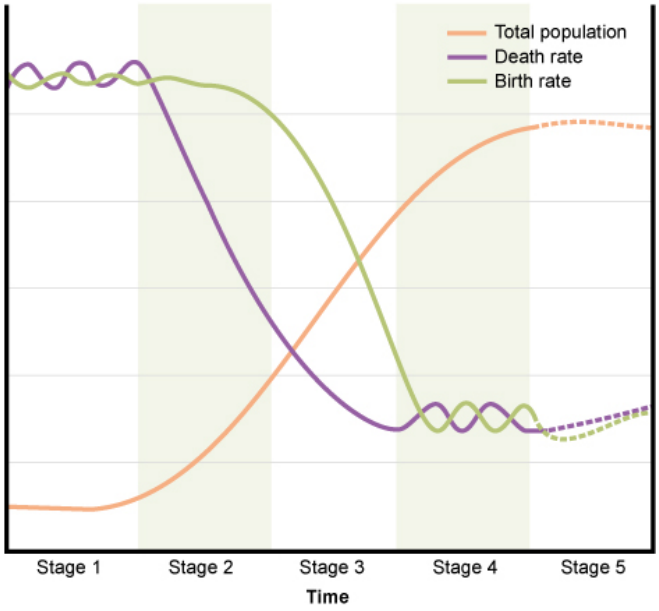

Fig. 1: Stages of Population Growth (Credits: The Open University)

predictions, particularly for mid-aged groups. However, deep learning models struggled with data variability at extreme age groups. The study shows the potential of LSTM networks in projection of mortality rate in a deterministic fashion, as was the primary source followed for this paper.

Raftery, et al developed a Bayesian hierarchical model for Total Fertility Rate (TFR) projections, which correlates to birth rate projections. This was done by decomposing fertility transitions into three phases: high fertility, transition, and low fertility, and utilizing a double logistic function to determine the outcome.

$$g(f,\theta) = \frac{d}{1+e^{-b_1(f-c_1)}} - \frac{d}{1+e^{-b_2(f-c_2)}} \tag{1}$$

where $\theta = (d, b_1, b_2, c_1, c_2)$

The model integrates global trends, while also utilizing country-specific data, offering probabilistic projections to account for uncertainty. The probabilistic element is introduced by Monte Carlo method, by introducing Gaussian noise to the output parameter with standard deviation increasing with time. These distribution parameters are based on the previous mean and standard deviation in the time series data. This error term becomes especially prominent in the stable region of birth rate.

$$\varepsilon_{c,t} \sim N(0, \sigma_t^2) \tag{2}$$

Validation showed the model outperformed deterministic methods, particularly in countries experiencing rapid fertility transitions. This approach was tried, however, a double logistic function is only applicable for a subset of countries and thus data required intensive pre-processing, which was out of the scope of the current approach. This method also allows for determination of confidence intervals of data.

Alkema, et al made long-term projections using autoregressive, Bayesian-hierarchial models by integrating climate and socioeconomic feedback into demographic models. The logistic model used by the previous paper was derived from these models.

Wilson, Tom et al examined multivariate models that include fertility, mortality, migration, and socioeconomic determinants such as education and healthcare access. Their findings demonstrated the impact of these factors on population transitions, emphasizing the importance of a holistic approach for effective policy planning.

In addition, various ensemble learning techniques were examined before choosing the LSTM Model for this project.

III. Dataset

A comprehensive dataset, which published by United Nations, containing data from 1950 to 2023 of various countries, including Year, Population (thousands), Births (thousands), Total Fertility Rate, Deaths (thousands), Life Expectancy at Birth (years), Net Number of Migrants (thousands) Migration Rate (per 1,000 people), was utilized for this project.

IV. Methodology

The size of the input data for a specific country is 74×8, which is a relatively small dataset to predicting population.

A. Preprocessing

To increase the size of the dataset, the data fields of Birth Rate, Death Rate and Migration Rate (which are obtained by dividing the number of births, the deaths and the net number of migrants in thousands, respectively, by the Population) are upsampled by a factor of 50 using quadratic interpolation. As the slope of the data is an important factor for LSTM, linear interpolation is not preferred. FFT-based interpolation could also be used in larger datasets (though not implemented here due to edge effects).

$$X(k) = \sum_{i=0}^{N-1} x_i e^{-j2\pi i/N} \tag{3}$$

$$Y(k) = X(k) \text{ padded with zeros, centred at} \frac{N}{2} \tag{4}$$

where, N is the number of data points and Y_i is the upsampled version of X_i.

In addition, an exponential moving average filter is used to filter out the high-frequency noise in the data, which may arise due to sudden temporary changes in birth or death rate, such as the Bengal Famine or the COVID-19 pandemic, as well as Government Policies. While a trade-off of accuracy being slightly lost in timelines of high slope exists, the general trend is captured, which is the more important part for LSTM.

$$x_n = \frac{1}{10} \sum_{i=n-9}^{n} x_i \tag{5}$$

By training a part of LSTM on a larger set of global data, we could capture certain global trends also, including the 3-phase curve of Birth Rate. A double logistic function

Layer (type)	Output Shape	Param #
lstm_42 (LSTM)	(None, 10, 100)	41,600
dropout_42 (Dropout)	(None, 10, 100)	0
lstm_43 (LSTM)	(None, 10, 120)	106,080
dropout_43 (Dropout)	(None, 10, 120)	0
lstm_44 (LSTM)	(None, 140)	146,160
dropout_44 (Dropout)	(None, 140)	0
dense_14 (Dense)	(None, 3)	423

Total params: 294,263 (1.12 MB)
Trainable params: 294,263 (1.12 MB)
Non-trainable params: 0 (0.00 B)

Fig. 2: Parameters metadata in the LSTM

also works for certain nations, though skipped in the final implementation.

LSTMs require sequential input in order to capture temporal dependencies. Here, sequences are prepared as overlapping sliding windows. Input sequences contain the last n time steps of demographic rates (e.g., 10 years of data), while the output labels represent the rates at the next time step. For instance, in Nigeria's case, if the sequence length is 10, the input is Birth, death, and migration rates from 2010–2019, and the output is these rates for 2020.

B. Model Architecture

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed to model sequential data. LSTMs effectively handle long-term dependencies by using gates (forget, input, and output) to manage memory.

In this code, the model architecture is as follows:

1) Input Layer: Accepts sequences of shape (sequence length, 3) (10 years of rates for 3 variables).
2) First LSTM layer with 100 units, processing sequences and outputting intermediate representations.
3) Second LSTM layer with 120 units, deepening the model's ability to learn complex dependencies.
4) Third LSTM layer with 140 units, producing a final representation for predictions.
5) Dropout Layers: Regularization layers with 20% dropout prevent overfitting by randomly deactivating neurons during training.
6) Output Layer: A dense layer with 3 units outputs predictions for birth, death, and migration rates.

A problem frequently encountered throughout the project was that of gradient explosion, often leading to floating-point precision loss at very large numbers and thus taking the actual value of birth and mortality rate after multiplication to very large numbers - either positive or negative. In other to avoid this, dropout layers were added for regularization, in addition to fine-tuning the hyperparameters (specifically, learning rate and number of epochs).

C. Training Loss

The training loss is computed using the least squares method.

$$L(\mathbf{x},\mathbf{y},n) = \sum_{i=0}^{n-1} \frac{1}{2}(y_i - f(x_i))^2 \tag{6}$$

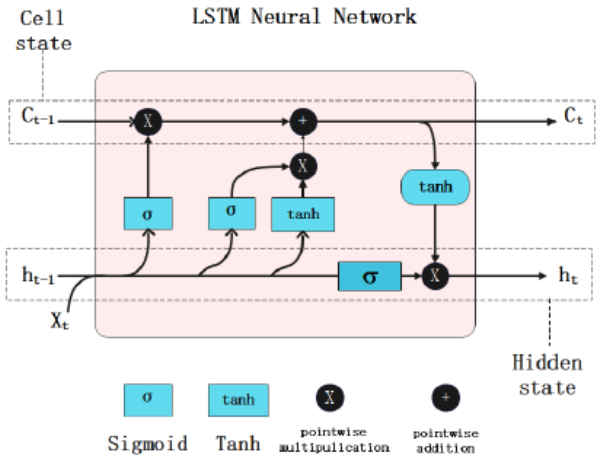

Fig. 3: LSTM Unit Cell (credits: Chen et al, 2023)

Epoch	Loss
1	5.8240×10^{-5}
2	7.1889×10^{-6}
3	6.5795×10^{-6}
4	5.9467×10^{-6}
5	5.9079×10^{-6}
6	5.7954×10^{-6}
7	5.7814×10^{-6}
8	5.5121×10^{-6}
9	5.4562×10^{-6}
10	5.3609×10^{-6}
11	5.4269×10^{-6}
12	5.2747×10^{-6}
13	5.0482×10^{-6}
14	4.9339×10^{-6}
15	4.6187×10^{-6}
16	4.5789×10^{-6}
17	4.3014×10^{-6}
18	4.2262×10^{-6}
19	3.6050×10^{-6}
20	3.4246×10^{-6}
21	2.8284×10^{-6}
22	2.3011×10^{-6}
23	1.7028×10^{-6}
24	1.1631×10^{-6}
25	8.1134×10^{-7}

TABLE I: Training loss at each epoch using the least squares method.

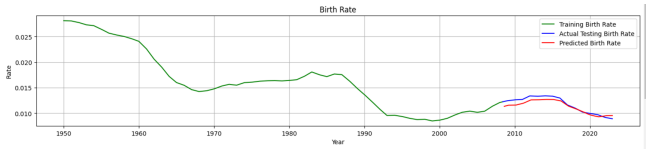

Fig. 4: Example of Migration Rate Prediction: Russian Federation. Sharp discontinuities between training and prediction are later manually removed by using a filter

D. Training and Testing

For initial training and testing of model, around 80% of the available data is utilized for training and 20% for testing.

E. Results - Case Studies

The countries Russia, United States of America and Nigeria as case studies. Russia is a nation with shrinking population, while Nigeria is rapidly growing. USA is a special case where the population is actually shrinking yet the migration rate keeps the population increasing.

The testing population graph of United States is shown below, closely matching the actual data. In reality, migrations increased to a greater extent due to geopolitical events during the Obama administration, thus giving a higher slope than predicted. This highlights the importance of using a probabilistic model, which was skipped in this implementation. Despite this, the deterministic model produced relatively good accuracy. In addition, the sudden slowing down of population growth due to COVID-19 pandemic is also difficult to predict by LSTM, as irregularities were filtered out from training data.

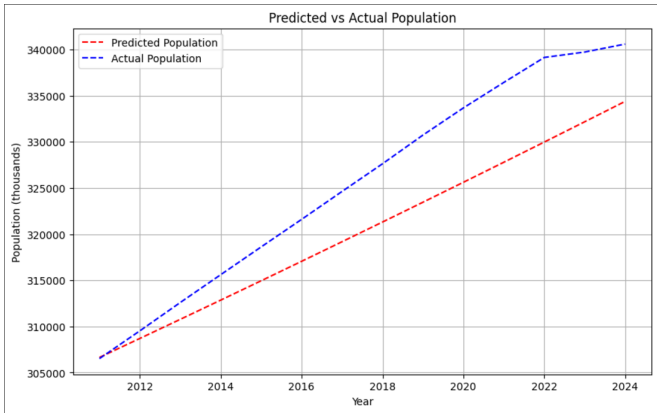

Fig. 5: Testing vs Actual Population Data in USA

The projections made for 2100 by this model closely matches with that of United Nations (around 420 million people, nearly constant rise).

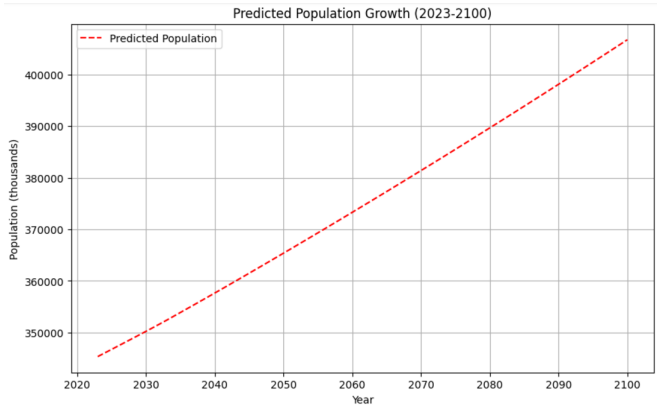

Fig. 6: Predicted Population Data in USA

Nigeria is a country which is still in early phase of its demographic transition, with high but slowly falling birth rates.

This model predicts a population of 1.17 billion for Nigeria by 2100, as compared to 790 million by United Nations, an error of over 40%. This is due to various socio-economic factors as well as relatively low life expectancy, which was also not taken into account. The usage of an autoencoder to correlate birth rate and death rate separated by a time period equal to the life expectancy could solve such errors in prediction.

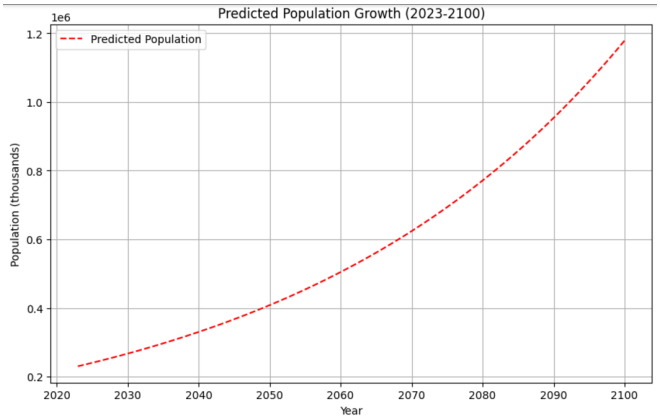

Fig. 7: Predicted Population Data in Nigeria

V. Conclusion and Future Work

This LSTM-based approach effectively models and forecasts demographic trends. By capturing temporal dependencies in birth, death, and migration rates, the methodology provides valuable insights into population dynamics. Applying these techniques to the USA, China, and Nigeria highlights their diverse demographic transitions and the potential for tailored policy interventions.

However, many aspects mentioned throughout this report wherever necessary, such as a probabilistic model using Monte Carlo framework, variational autoencoders, socio-economic factors correlating to important parameters (which can be correlated after dimensionality reduction), were missed out in my implementation due to time constraint. However, this gave rise to direct learnings in terms of utilising of techniques to increase the effectiveness of models, the utilization of preprocessing techniques, multi-task learning to strike a balance between uniqueness and scalability of models, parameter estimation in a deterministic setting, and model selection, as well as indirect learnings (from literature review) in terms of usage of Bayesian Prediction Models for time series data, matrix methods for formulating projections, parameter estimation in a probabilistic setting, noise modelling and practical exposure to large datasets, as well as a perspective into expanding on and combining concepts from existing papers to create something original.

VI. References

1) Chen, Yuan and Khaliq, Abdul Q. M. (2023). Mortality Rates Forecasting with Data-Driven LSTM, Bi-LSTM, and GRU: the United States Case Study. Journal of Mortality Studies, 57(3), 12-34.

2) Raftery, Adrian E. and Ševčíková, Hana (2023). Probabilistic Population Projection: Short Term to Very-Long Term. International Journal of Forecasting, 39(1), 73-97.
3) Alkema, Leontine and Raftery, Adrian E. and Gerland, Patrick and Clark, Samuel J. and Pelletier, François and Buettner, Thomas and Heilig, Gerhard K. (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. Demography, 48(3), 815-839.
4) Vollset, Stein Emil and Others (2020). Long-term Population Projections and Economic Implications. Population Studies, 57(4), 112-139.
5) Wilson, Tom and Rees, Philip Howell (2004). Recent Developments in Population Projection Methodology. Global Demography, 59(2), 89-105.